

**Regression Modeling for Competing Risks Based on Pseudo-Observations, with  
Application to Breast Cancer Study**

by

**Shangzhen Chen**

BS, University of Missouri, 2017

Submitted to the Graduate Faculty of

Department of Biostatistics

Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Master of Science

University of Pittsburgh

2019

UNIVERSITY OF PITTSBURGH  
GRADUATE SCHOOL OF PUBLIC HEALTH

This thesis was presented

by

**Shangzhen Chen**

It was defended on

June 5, 2019

and approved by

Ada O Youk, PhD

Associate Professor, Department of Biostatistics  
Graduate School of Public Health  
University of Pittsburgh

Douglas Landsittel, PhD

Professor, Department of Biomedical Informatics  
Professor, Department of Biostatistics  
School of Medicine  
University of Pittsburgh

**Thesis Advisor:** Jong H Jeong, PhD

Professor, Department of Biostatistics  
Graduate School of Public Health  
University of Pittsburgh

Copyright © by Shangzhen Chen

2019

# **Regression Modeling for Competing Risks Based on Pseudo-Observations, with Application to Breast Cancer Study**

Shangzhen Chen, MS

University of Pittsburgh, 2019

## **Abstract**

In medical research, a patient might experience a failure due to different causes, where each cause would be considered a competing risk. The standard method to analyze the type of data that is used most frequently in medical research is the cause-specific hazard regression analysis based on the Cox proportional hazards model or the cumulative incidence function-based method. The Cox proportional hazards model is used to model the cause-specific event rates and treat the other type of events as independent censoring. The new methods which have been proposed directly assess the effect of covariates on the cumulative incidence functions, which is the pseudo-observations approach proposed by Andersen and Klein in 2007. The scheme of pseudo-observations approach is to (1) choose some fixed time points that are equally spaced on the event scale, (2) calculate pseudo-observations for each individual at those fixed time points using the jackknife technique, and (3) fit a generalized linear model with GEE method based on the conditional cumulative incidence function using the pseudo-observations.

We applied the pseudo-observation approach to a breast cancer study. The goal of the study was to assess the effect of the covariates on the cumulative incidence function. The results showed that nodal status and tumor size are positively related to cumulative incidence of death following breast cancer recurrence and that age has a negative relationship with the cumulative incidence of death following breast cancer recurrence. In addition, nodal status and tumor size are not

significantly associated with death due to causes other than breast cancer. Age is positively related to death not due to breast cancer.

**Public health significance:** Because the method explored and applied here is a readily accessible procedure for censored time-to-event data, providing straightforward interpretation of the effect of the predictors on the cumulative incidence function, dissemination of the method through a real world example would help the researchers, stakeholders, and patients in public health understand the usefulness of the method and grasp the interpretation of the analysis for competing risks data.

## Table of Contents

<b>1.0 Introduction .....</b>	<b>1</b>
<b>2.0 Methodology .....</b>	<b>5</b>
<b>2.1 Cumulative Incidence Function .....</b>	<b>5</b>
<b>2.2 Jack-knife statistic and Pseudo-Observations .....</b>	<b>6</b>
<b>2.3 Generalized Linear Model with Generalized estimating Equation (GEE) .....</b>	<b>7</b>
<b>3.0 Result.....</b>	<b>10</b>
<b>3.1 Bone Marrow Transplant study.....</b>	<b>10</b>
<b>3.2 NSABP B-04 Study .....</b>	<b>11</b>
<b>4.0 Discussion.....</b>	<b>21</b>
<b>Appendix A R code .....</b>	<b>23</b>
<b>Bibliography .....</b>	<b>37</b>

## **List of Tables**

Table 1 Complementary Log-Log Model for Relapse.....	11
Table 2 Cox Proportional Hazard Model for Death Following Breast Cancer Recurrence .....	13
Table 3 Test of Proportional Hazard in Model for Death following Breast Cancer Recurrence..	13
Table 4 Cox Proportional Hazard Model for Death Not Related To Breast Cancer .....	14
Table 5 Test of Proportional Hazard in Model for Death Not Related To Breast Cancer.....	15
Table 6 Complementary Log-Log Model and Logistic Model for Breast Cancer-Related Death	17
Table 7 Complementary Log-Log Model and Logistic Model for Death Not Related to Breast Cancer .....	19

## **List of Figures**

Figure 1 Cumulative Incidence Curve of type 1 and type 2 events .....	12
Figure 2 Schonefeld Residual Plot for Model of Death following Breast Cancer Recurrence ....	14
Figure 3 Schonefeld Residual Plot for Model of Death not Related To Breast Cancer .....	15



## 1.0 Introduction

Survival data measure the time from a reference time point to an event time point [14]. If there are no incomplete observations in survival data, the standard methods for the continuous outcome or binary outcome can be used for analysis [2]. However, since survival data is often subject to right censoring, where the subject leaves the cohort before experiencing the event, survival methods must account for censoring through parametric or non-parametric estimation, or inference, either through unadjusted survival distributions or multivariate regression analysis [2].

For censored survival data, one way of adopting the standard method is to replace a summary measure of interest as a function of time, say  $f(x)$ , by their pseudo-observations [2]. The estimator  $\hat{\theta}$  is the expected value of  $f(x)$  exists no matter whether the data are complete or not [2]. If the data are complete, the expected value can be estimated by  $\frac{\sum_i f(x_i)}{n}$ . If the data are not complete, an estimate of the expected value is also available such as Kaplan-Meier estimator for survival function [2]. The pseudo-observation is defined as  $\hat{\theta}_i = n\hat{\theta} - (n-1)\hat{\theta}^{-i}$ ,  $i = 1, \dots, n$ , where  $\hat{\theta}^{-i}$  is an estimator based on sample size of  $(n-1)$  after the  $i^{\text{th}}$  observation is deleted from the data set. The pseudo-observation  $\hat{\theta}_i$  is used for all individuals not only for the unobserved individuals (censored objects) [2].

Pseudo-observations can be applied to several areas in survival analysis such as the survival function, the restricted mean survival time and transition or state occupation probabilities in multi-state models [2]. Competing risks model is a special case of multi-state models. It only has an initial state and some exclusive causes of final state, whereas multi-state model is the model which

not only contains initial state and final states but also has intermediate events [8]. In this current work, we apply the pseudo observation method to competing risks regression.

Competing risks happen when the event of interest is not observed due to another type of event that precedes it. For example, in a cancer study, researchers want to investigate the time from the beginning of a treatment to tumor-related death. In this situation, the patients that die due to the other causes become competing events. There are several ways of analyzing survival data with competing risks such as cause-specific hazard regression, subdistribution hazard regression, mixture models, vertical modelling and regression modeling of the competing risks data based on pseudo-observations [8]. This thesis focuses on regression modeling of the competing risks data based on pseudo-observations. Competing risks probabilities can be summarized by the cumulative incidence function [11]. This function defines the cumulative probability of subject who fail from the cause  $j$ .

$$C_j(t) = \Pr[T \leq t, \epsilon = j] = \int_0^t h_j(u) S(u-) du \quad (1)$$

$$\text{Where } S(u-) = P(T \geq u) \text{ and } h_j(u) = \lim_{\Delta u \rightarrow 0} \frac{1}{\Delta u} P(u \leq T < u + \Delta u, \epsilon = j | T \geq u)$$

In this thesis, we formulate a regression model of breast cancer data based on pseudo observations. Breast cancer is the most common cancer in women [9]. About one in eight to ten women will get breast cancer in their lifetime [9]. There are five stages of breast cancer, and doctors use TNM system to describe the stages. TNM represents Tumor (T), Node (N), and Metastasis (M), where tumor means the size and location of a tumor [6]. Tumor size can help doctors figure out the prognosis such as the likely outcome of the disease, decide the best treatment option, and determine whether a certain clinical trial may fit for the patient. Node means whether the tumor spread to the lymph nodes [6]. There are two lymph node status; negative and positive. Lymph node-negative means that the tumor does not spread to axillary lymph nodes [12]. Lymph

node-positive means that the axillary lymph nodes contain the cancer [12]. Lymph node status is also a reference for doctors to determine the effect of the prognosis [12]. The prognosis is better if the tumor does not spread to the lymph nodes. Metastasis means that the cancer has spread to the other part of body. Stage 0 describes disease where the tumor has not spread to the surrounding tissue of the breast. In stage 4, the tumor has spread to the other organs. There are several ways to treat breast cancer. Stage 0 is the least severe status of breast cancer, and stage 4 is the most severe status of breast cancer. The treatment that doctor chooses depends on the types of breast cancer and how far it has spread. People with breast cancer often get a combination of treatments. The treatments include surgery, chemotherapy, hormonal therapy, biological therapy and radiation therapy. The patients in the study we are analyzing in this thesis were treated with surgery and radiation therapy.

Data for this study were taken from the B-04 phase III randomized clinical trial conducted by the National Surgical Adjuvant Breast and Bowel Project (NSABP). Participants were randomized to one of two types of surgical treatments. One was traditional radical mastectomy; the other was less aggressive total mastectomy. Traditional radical mastectomy is also called Halsted mastectomy, which is no longer a common procedure unless patients have severe breast cancer which has invaded muscle under the breast tissue. A radical mastectomy removes all of the breast tissue with the tumor and removes all of the lymph nodes under the arm and muscle which lies under the breast, whereas total mastectomy only removes the breast tissue with the tumor [16]. The B-04 study compared the two treatments on survival of patients. The women were separated into two groups. The women in the first group were with negative axillary nodes, and the women in the second group were with positive axillary nodes. The patients in one of groups were treated with traditional radical mastectomy or less aggressive total mastectomy randomly [4]. The study

lasted 30 years, and about 30% of patients with negative node and 20% of patients with positive node were censored. Patients could be censored, died following breast cancer recurrence, or died due to other reasons. The study showed that the less aggressive mastectomy is better than the traditional aggressive surgery [4].

In this thesis, we will formulate generalized linear model with GEE based on pseudo-observations to assess the effect of covariates on cumulative incidence function. In the method section, we will explain the concepts and the procedure of regression modeling of competing risks based on pseudo-observations in detail. In the result section, we will apply the pseudo-observations method to NSABP B-04 study and find the relationship between the cumulative incidence and the covariates.

## 2.0 Methodology

In this section, we present an approach to formulate a regression model based on the cumulative incidence function. The reasons why we apply pseudo-observations methods to cumulative incidence function are that (1) the pseudo-observations method for censored survival data makes the regression modeling more flexible, allowing for using different link functions, (2) people can use the standard generalized linear model to estimate the parameters and get the relationship between covariates and cumulative incidence directly, (3) compared to cause-specific hazard regression or subdistribution hazard regression analysis, the generalized linear model with GEE based on the pseudo-observations does not require strict assumption such as the proportional hazard assumption.

### 2.1 Cumulative Incidence Function

The cumulative incidence function is defined as the probability of a particular event which is due to cause  $j$  occurring before a given time [10]. The cumulative incidence function is a nondecreasing function of time with  $C_j(\infty) = \Pr[\epsilon = j]$ . [10] The estimated cumulative incidence function for cause  $j$  is [2,10]:

$$\hat{C}_j(t) = \Pr[T \leq t, \epsilon = j] = \int_0^t \hat{S}(u-) d\hat{A}_j(u) = \int_0^t \prod_{T_i < u} \left(1 - \frac{\sum_{h=1}^K dN_h(T_i)}{Y(T_i)}\right) \frac{dN_j(u)}{Y(u)} \quad (2)$$

In this formula,  $\hat{A}_j(u) = \int_0^u \frac{dN_j(u)}{Y(u)}$  is the Nelson-Aalen estimator for the cumulative cause-specific hazard  $A_j(t)$  for cause  $j$  failure.  $N_j(u)$  is a counting process  $N_j(t) = \sum_i I(\tilde{X}_i \leq t, D_i = j)$ ,  $j = 1, 2$ , where  $\tilde{X}_i$  is the right censored or failure time and  $D_i$  is the event the individual experiences.  $\prod_{T_i < u} (1 - \frac{\sum_{h=1}^K dN_h(T_i)}{Y(T_i)})$  is the Kaplan-Meier estimate before time  $u$ .

The formula can also be written as [9]:

$$\hat{C}_j(t) = \sum_{t_h \leq t} \frac{d_{hk}}{Y_h} \prod_{t_i < t_h} \left[ \frac{Y_i - (d_{1i} + d_{2i})}{Y_i} \right], i = 1, \dots, n, j = 1, 2, h = 1, \dots, M \quad (3)$$

Where  $t_h$  is the fixed time points which people choose,  $t_i$  is the time where one of events happens,  $Y_h$  is the number at risk at time  $t_h$ ,  $d_{1h}$  and  $d_{2h}$  are the number of type 1 and type 2 events at time  $t_h$ , and  $\prod_{t_i < t_h} \left[ \frac{Y_i - (d_{1i} + d_{2i})}{Y_i} \right]$  is the Kaplan Meier estimator of the survival function.

## 2.2 Jack-knife statistic and Pseudo-Observations

The pseudo-observations are from a jackknife technique [3]. The Jackknife procedure is a method of resampling [12]. Each jackknife sample is selected from the original data and deleted one observation from the set [12]. So, the  $i^{\text{th}}$  Jackknife sample vector is like:

$$\mathbf{X}_{[i]} = \{X_1, X_2, \dots, X_{i-1}, X_{i+1}, \dots, X_{n-1}, X_n\}.$$

The Jackknife procedure can be used to generate the “pseudo-values”. The pseudo values are treated as the independent random variables. The  $i^{\text{th}}$  pseudo-values are defined as:

$$\hat{\theta}_i = n\hat{\theta} - (n-1)\hat{\theta}^{-i}. \quad (4)$$

where  $\theta = E[f(X)]$ ,  $\hat{\theta}$  is an unbiased estimator (or approximately unbiased) estimator of  $\theta$ , and  $\hat{\theta}^{-i}$  is the estimator whose sample size is  $n-1$  due to eliminating the  $i^{\text{th}}$  observation from the data. A Jackknife pseudo value can be viewed as a biased-corrected estimator, which correct the bias  $\widehat{bias}_{jack} = (n - 1)(\hat{\theta}^{-i} - \hat{\theta})$  [13].

The pseudo observations applied in competing risks in this thesis are from the cumulative incidence function, implying that our  $f(X)$  will be the cumulative incidence function. For fixed time points  $(\tau_1, \tau_2, \tau_3, \dots, \tau_M)$ , there will be  $M$  cumulative incidence estimates for each observation. The estimated cumulative incidence function at each time point is calculated based on cumulative incidence function  $\hat{C}_j(\tau_h)$  which is derived from the complete data set and the cumulative incidence function  $\hat{C}_j^{(i)}(\tau_h)$  which is derived from sample size  $n-1$  [10]. The pseudo-values for the  $i^{\text{th}}$  observation at time  $\tau_h$  is defined as:

$$\hat{\theta}_{ih} = n\hat{C}_j(\tau_h) - (n - 1)\hat{C}_j^{(i)}(\tau_h), i = 1, \dots, n, j = 1, \dots, k, h = 1, \dots, M. \quad (5)$$

where  $i$  represents observations,  $j$  represents causes, and  $h$  represents time points.

If there is no censoring, there are  $n\hat{C}_j(\tau_h)$  events of type  $j$  occurring up to time  $t$ , and  $\hat{\theta}_i$ 's are independent. When there is censoring,  $\hat{\theta}_i$ 's are close to the number of the type  $j$  events and are approximately independent [11]. Because the pseudo values are independent between different individuals, we can use the results of generalized linear model to model the effects of covariates on the outcomes (cumulative incidence function) [11].

### 2.3 Generalized Linear Model with Generalized estimating Equation (GEE)

We assume a generalized linear model

$$g(\theta_{ih}) = \alpha_h + \gamma^T Z_{ih} = \beta^T Z_{ih}, i = 1, \dots, n, h = 1, \dots, M.$$

where  $g(\cdot)$  is the link function.  $\theta_{ih}$  is the pseudo-observations.  $\alpha_h$  is the intercept when covariates are equal to zero.  $Z_{ih}$  represent covariates.

Two link functions for the cumulative incidence are considered for this model. One is complementary log-log function on  $1-x$ , i.e.  $g(x) = \log(-\log(1-x))$  and the other is logit function on  $x$ , i.e.  $g(x) = \log \frac{x}{1-x}$ . When the complementary log-log link function is applied to the survival function, it mimics the proportional hazard on the subdistribution hazards function [7].

The regression model with the cumulative incidence function for logistic link function for specific cause  $j$  is

$$\text{logit}\left(C_{ij}(t_h|Z)\right) = \log\left(\frac{C_{ij}(t_h|Z)}{1-C_{ij}(t_h|Z)}\right) = \beta^T Z_{ijh}, i = 1, \dots, n, h = 1, \dots, M. \quad (6)$$

The regression model with the cumulative incidence function for complementary log-log link function for specific cause  $j$  is

$$\log\left(-\log\left(1 - C_{ij}(t_h|Z)\right)\right) = \beta^T Z_{ijh}, i = 1, \dots, n, h = 1, \dots, M. \quad (7)$$

Since the pseudo observations at different time points (for an individual) are correlated, we use the generalized estimating equation (GEE) method to estimate the regression parameters  $\beta$  and variance. The GEE approach, as introduced by Liang and Zeger (1986), models the marginal mean structure. To solve the estimated  $\beta$ , we need to solve the equation

$$U(\beta) = \sum_i U_i(\beta) = \sum_i \left( \frac{d}{d\beta} g^{-1}(\beta^T Z_{ih}) \right)^T V_i^{-1}(\beta) (\hat{\theta}_i - g^{-1}(\beta^T Z_{ih})) = 0 \quad (8)$$

where  $g^{-1}(\beta^T Z_{ih}) = \theta_{ih}$ , and  $V_i(\beta)$  is the working covariance matrix.

To estimate the variance of  $\hat{\beta}$ , a sandwich estimator is calculated as

$$\widehat{\text{var}}(\hat{\beta}) = I(\hat{\beta})^{-1} \widehat{\text{var}}(U(\beta)) (I(\hat{\beta})^{-1}).$$

where



$$I(\beta) = \sum_i \left( \frac{dg^{-1}(\beta^T Z_i)}{d\beta} \right)^T V_i^{-1} \left( \frac{dg^{-1}(\beta^T Z_i)}{d\beta} \right), \text{ and}$$

$$\widehat{var}(U(\beta)) = \sum_i U_i(\hat{\beta})^T U_i(\hat{\beta}).$$

In summary, we presented an approach to formulate a regression model based on the cumulative incidence function. Suppose we have data with  $n$  individuals, and each observation consists of four components; observation time ( $T_i$ ), censoring indicator ( $\delta_i$ ), type of event ( $\varepsilon_i$ ) and covariates ( $X_i$ ). To formulate a regression model, first we choose the number of time points which are equally spaced on event scale, second we calculate two kinds of cumulative incidence function of the survival data base on full data and “leave-one-out” data, third we create the pseudo observations, and finally we apply the generalized linear model with GEE based on the pseudo-observations.

### **3.0 Result**

#### **3.1 Bone Marrow Transplant study**

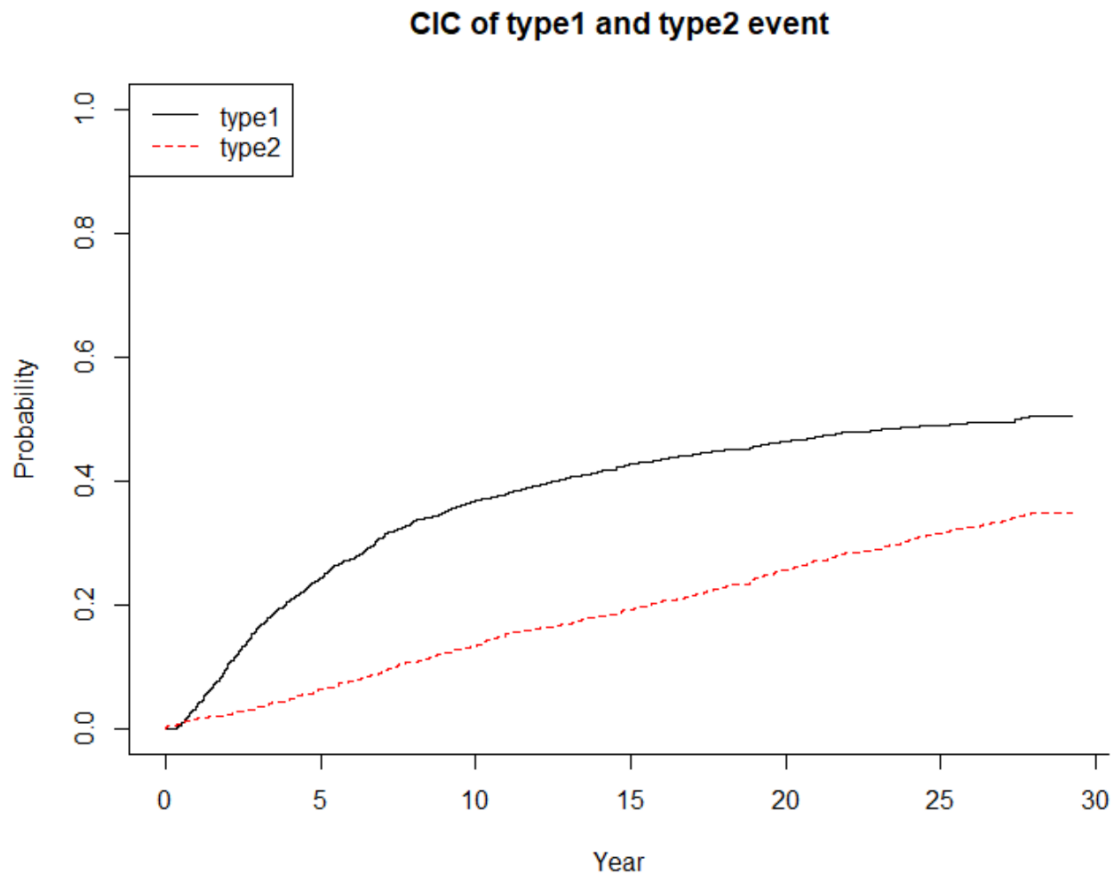
First, we reproduced the analysis results of Bone Marrow Transplant Study to verify our codes that will be used to analyze NSABP B-04 study. Andersen and Klein (2007) applied the pseudo-observations method in survival analysis. They formulated regression models for parameters like the survival function, the restricted mean survival time, and the cumulative incidence function. In this thesis, we focus on regression modeling for competing risk data based on pseudo observation methods. Andersen and Klein (2007) applied the method on bone marrow transplant data which was in the R package KMsurv. The data set contained 137 observations. They focused on investigating the cumulative incidence of relapse. They fit a generalized linear model with the GEE method based on the pseudo values of the cumulative incidence of relapse. They chose the model with complementary log-log link and independent working covariance. Complementary log-log transformation was applied to pseudo values, and types of disease, age, types of French-American-British (FAB) which is a classification system of acute myeloblastic leukemia (AML) and different time points as covariates. We obtained the same results as in Andersen and Klein (2007), which was shown in the Table 1.

**Table 1 Complementary Log-Log Model for Relapse**

Relapse					
Link Function = complementary log-log					
Variable		Mean	SD	Z	p-value
Time Points	50	-3.55	0.85	-4.17	<0.001
	105	-2.54	0.67	-3.76	0.0002
	170	-2.07	0.66	-3.14	0.002
	280	-1.73	0.64	-2.68	0.007
	530	-1.43	0.63	-2.27	0.02
AML Low Risk		-1.77	0.67	-2.62	0.0089
AML High Risk		-0.24	0.59	-0.41	0.68
FAB		1.13	0.53	2.16	0.031
Age		0.01	0.02	0.63	0.53

### 3.2 NSABP B-04 Study

The NSABP B-04 study contained 1,599 patients, with data on patient ID, survival time, type of event, nodal status, age and tumor size, etc. 48% of the entire patients died following breast cancer recurrence (type 1), 30% died for reasons unrelated to breast cancer (type 2), and 23% were censored. 1038 (65%) patients in lymph-node negative status (nodal status 0) and 561 (35%) patients in lymph-node positive status (nodal status 1). Figure 1 shows a greater cumulative incidence for death following breast cancer recurrence versus death unrelated to breast cancer.



**Figure 1 Cumulative Incidence Curve of type 1 and type 2 events**

First, we estimated the effects of the covariates in the Cox model on the cause-specific hazard function and checked whether the Cox model fit the proportional hazards assumption. The Cox model for the cause-specific hazard function is the common model used to examine the effects of covariates on specific cause of failure. For the type 1 events, the p-values from the Cox proportional hazards model in Table 2 show that nodal status and tumor size are positively associated with death following breast cancer recurrence. In Table 3, the p-values indicate nonproportional hazards in all the covariates and hence globally, and Figure 2 also indicates that the log-hazard ratio changes over time. The Cox proportional hazards model for type 1 events violates the proportionality assumptions. In general, there are two ways to accommodate the non-

proportional hazards. One is to generate the stratified proportional hazards model; the other is to fit a model with time-dependent covariates using interaction terms. For type 2 events, the p-values of Cox proportional hazards model in Table 4 show that the hazard rate of death not due to breast cancer is higher when the patients are older. The Cox proportional hazards model for type 2 events does not violate the proportional hazard assumptions at the significance level of 0.05, and the lines in Figure 3 seems flat.

**Table 2 Cox Proportional Hazard Model for Death Following Breast Cancer Recurrence**

Type 1					
Model: Cox Proportional Hazard Model					
		coefficient	SE (coefficient)	Z	p-value
Nodal Status	1	0.56	0.07	7.70	<0.0001
Age		-0.10	0.32	-0.33	0.74
Tumor Size		0.82	0.14	6.00	<0.0001

**Table 3 Test of Proportional Hazard in Model for Death following Breast Cancer Recurrence**

Type 1				
Check proportional Hazard Assumption				
		rho	chisq	p-value
Nodal Status	1	-0.10	8.22	$4.14 \times 10^{-3}$
Age		0.23	44.15	<0.0001
Tumor Size		-0.09	3.52	$6.06 \times 10^{-2}$
Global		NA	62.30	<0.0001

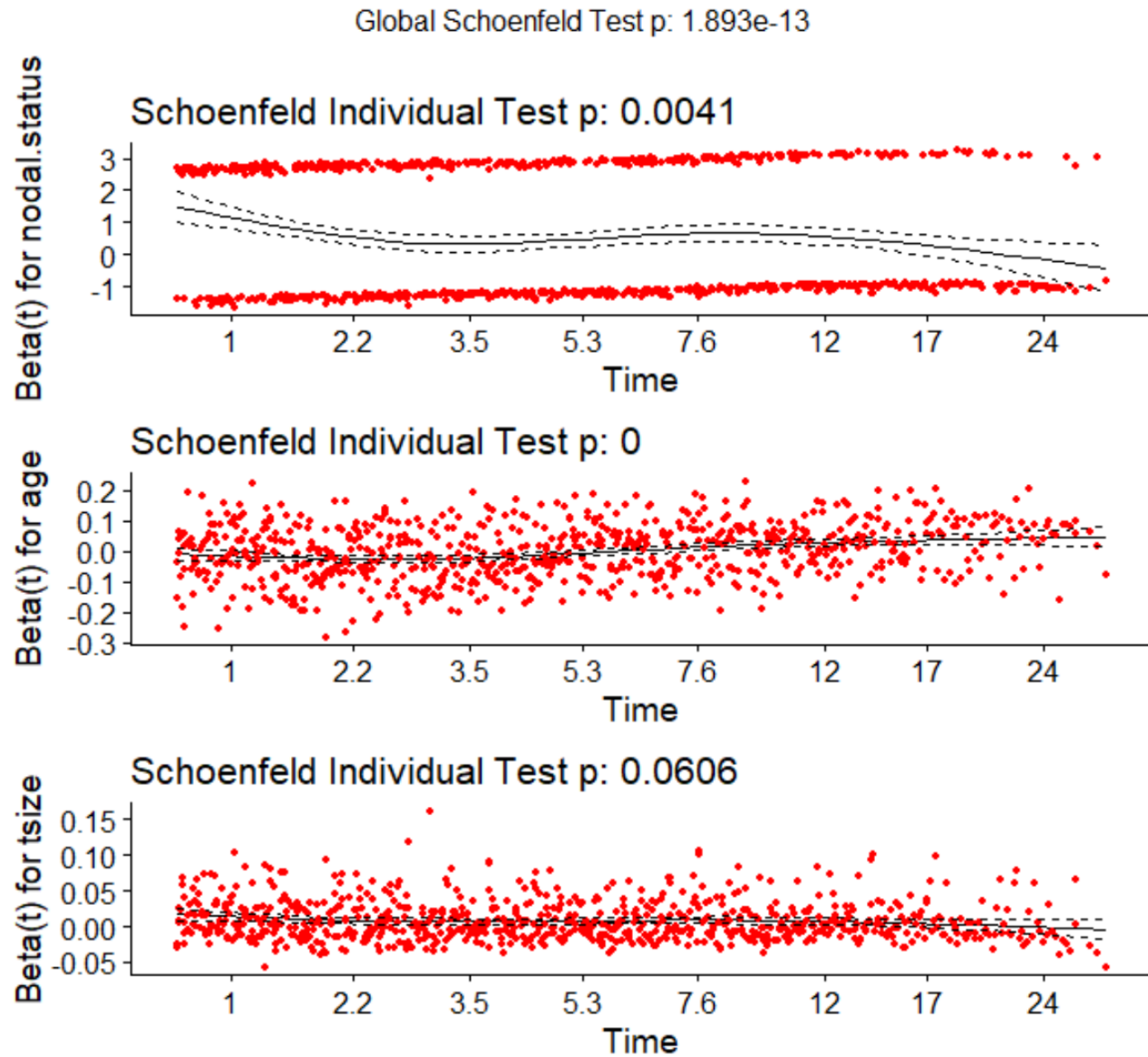


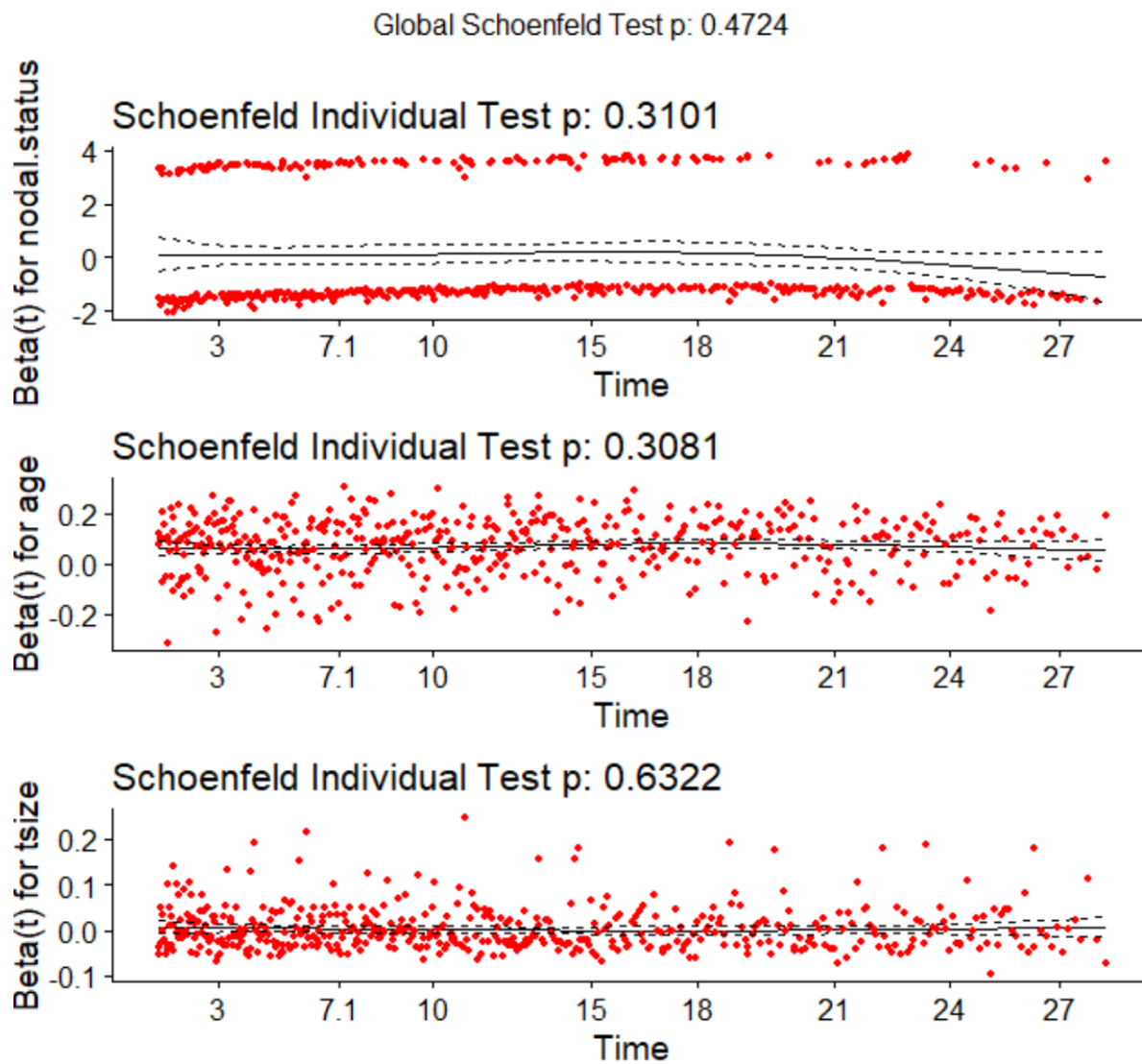
Figure 2 Schonefeld Residual Plot for Model of Death following Breast Cancer Recurrence

Table 4 Cox Proportional Hazard Model for Death Not Related To Breast Cancer

Type 2					
Model: Cox Proportional Hazard Model					
		coefficient	SE (coefficient)	Z	p-value
Nodal Status	1	0.06	0.1	0.58	0.57
Age		6.73	0.46	14.56	<0.0001
Tumor Size		0.29	0.23	1.25	0.21

**Table 5 Test of Proportional Hazard in Model for Death Not Related To Breast Cancer**

Type 2				
Check proportional Hazard Assumption				
		rho	chisq	p-value
Nodal Status	1	-0.05	1.03	0.31
Age		0.04	1.04	0.31
Tumor Size		-0.02	0.23	0.63
Global		NA	2.52	0.47



**Figure 3 Schonefeld Residual Plot for Model of Death not Related To Breast Cancer**

Second, we assessed the effects of covariates on the cumulative incidence function by formulating generalized linear model with GEE based on pseudo-observations. First, we chose the different time points. The strategy for us to choose the time points is the same as how Andersen and Klein did. The strategy is that we made sure that the number of events would be evenly distributed across time intervals.

To examine whether the number of time points and the link function influence the effect of covariate estimation, we varied the number of time points (12, 6 and 4), and used two different link functions (logit link and complementary log-log link functions).

Table 6 shows that results of the method for death following breast cancer recurrence based on 12, 6 and 4 time points and the independence working covariance models under both link functions. The number of the time points and link function does not influence the relationship between covariates and the outcome. According to the p-values of table 6, there is association between nodal status, tumor size, age and cumulative incidence. In the models with complementary log-log link, the positive value of mean ( $\beta$ ) for the covariates represents higher cumulative incidence following breast cancer recurrence for patients. As we expected, prognosis depends on tumor size and nodal status; larger tumor size and positive nodal status cause the worse prognosis. Furthermore, younger women at diagnosis are more likely to get aggressive tumors and have a higher chance to recur both in locoregional tumor and distant sites [5]. All the models with complementary log-log link reflect the relationship between the cumulative incidence and covariates. The patients with positive nodal status have higher a chance of death following breast cancer recurrence compared to the patients with negative nodal status; the patients with larger tumor size have a higher chance of breast cancer-related death, and the younger patients have a higher chance of breast cancer-related death. The models with logistic link show similar results.



The interpretation of the models with the logistic link is different with the model with complementary log-log link. We interpret  $\exp(\beta)$  as the odds ratio between the current category and baseline category. For example, with 12 time points the odds of positive nodal status is 2.08 times the odds of the negative nodal status, after holding the other covariates constant. There is a 1.5% increase in the odds of breast cancer-related death for one-unit increase in tumor size while a 1.7% decrease in the odds of breast cancer-related death for one-unit increase in age. In summary, the patients with positive nodal status, larger tumor size and younger in age have higher odds of breast cancer-related death.

**Table 6 Complementary Log-Log Model and Logistic Model for Breast Cancer-Related Death**

Type1										
Time Points = 12										
	Link Function = complementary log-log						Link Function = logistic			
Variables		mean	SD	Z	p-value		mean	SD	Z	p-value
Time Points	1.14	-2.96	0.23	12.88	<0.0001		-3.01	0.28	-10.64	<0.0001
	2.02	-2.08	0.21	-9.83	<0.0001		-2.06	0.27	-7.75	<0.0001
	2.87	-1.66	0.21	-8.05	<0.0001		-1.59	0.26	-6.09	<0.0001
	3.91	-1.36	0.21	-6.62	<0.0001		-1.25	0.26	-4.78	<0.0001
	5.1	-1.14	0.2	-5.6	<0.0001		-1	0.26	-3.82	0.0001
	6.5	-0.97	0.2	-4.77	<0.0001		-0.78	0.26	-2.99	0.003
	7.93	-0.8	0.2	-3.96	0.0001		-0.57	0.26	-2.19	0.03
	9.96	-0.67	0.2	-3.32	0.0009		-0.4	0.26	-1.56	0.12
	12.51	-0.56	0.2	-2.8	0.005		-0.26	0.26	-1.01	0.31
	15.66	-0.45	0.2	-2.27	0.02		-0.12	0.26	-0.46	0.64
	19.68	-0.36	0.2	-1.8	0.07		0.01	0.26	0.04	0.97
	24.84	-0.27	0.2	-1.35	0.18		0.14	0.26	0.53	0.6
Nodal Status	1	0.58	0.08	7.54	<0.0001		0.73	0.1	7.38	<0.0001
Tumor Size		0.01	0.002	5.71	<0.0001		0.015	0.003	5.73	<0.0001
Age		-0.01	0.003	-4.1	<0.0001		-0.017	0.004	-4.07	<0.0001
Time Points = 6										
	Link Function = complementary log-log						Link Function = logistic			
Variables		Mean	SD	Z	p-value		Mean	SD	Z	p-value
Time Points	2.02	-2.09	0.21	-9.91	<0.0001		-2.07	0.27	-7.78	<0.0001
	3.91	-1.37	0.2	-6.69	<0.0001		-1.26	0.26	-4.82	<0.0001

<b>Table 6 Continued</b>										
	6.5	-0.98	0.2	-4.85	<0.0001		-0.79	0.26	-3.04	0.002
	9.96	-0.68	0.2	-3.39	0.0007		-0.41	0.26	-1.61	0.11
	15.66	-0.47	0.2	-2.35	0.019		-0.13	0.26	-0.52	0.6
	24.84	-0.28	0.2	-1.42	0.15		0.12	0.26	0.47	0.64
Nodal Status	1	0.56	0.08	7.43	<0.0001		0.72	0.1	7.28	<0.0001
Tumor Size		0.01	0.002	5.72	<0.0001		0.01	0.003	5.72	<0.0001
Age		-0.01	0.003	-4	0.0001		-0.02	0.004	-3.99	0.0001
Time Points = 4										
	Link Function = complementary log-log						Link Function = logistic			
Variables		Mean	SD	Z	p-value		Mean	SD	Z	p-value
Time Points	2.87	-1.68	0.2	-8.24	<0.0001		-1.63	0.26	-6.21	<0.0001
	6.5	-0.99	0.2	-4.96	<0.0001		-0.82	0.26	-3.15	0.002
	12.5	-0.59	0.2	-2.97	0.003		-0.31	0.26	-1.19	0.23
	24.84	-0.3	0.2	-1.53	0.1		0.09	0.26	0.34	0.73
Nodal Status	1	0.55	0.08	7.26	<0.0001		0.71	0.1	7.12	<0.0001
Tumor Size		0.01	0.002	5.55	<0.0001		0.91	0.003	5.6	<0.0001
Age		-0.01	0.003	-3.84	0.0001		-0.02	0.004	-3.81	0.0001

Table 7 shows that the results for death not related to breast cancer with 12, 6 and 4 time points and independence working covariance. The results show that the result of the regression model does not change when the number of time points are 12 or 6. However, nodal status is marginally significant in the model with complementary log-log link function and the model with logistic link function when the number of time points is equal to 4. Because the number of the parameters we estimated in the models with 4 time points is less than the number of the parameters we estimated in the models with 12 or 6 time points, the degrees of freedom increase. So, the p-values of the models with 4 time points are smaller than that of the model with 12 or 6 time points. The models with the complementary log-log link function and ones with logistic link function show that the nodal status and tumor size are associated with death unrelated to breast cancer. Age is significantly associated with the death not due to breast cancer. Although the tumor size and nodal status are related to death following breast cancer recurrence, they are not significantly

related to death not due to breast cancer. The results of the regression analysis for death not due to breast cancer with complementary log-log link function shows that the older patients have higher chance of death. The result of the regression analysis for death not due to breast cancer with the logistic link function shows that the older patients have higher odds of death not due to breast cancer compared to the younger patients.

**Table 7 Complementary Log-Log Model and Logistic Model for Death Not Related to Breast Cancer**

Type2										
Time Points = 12										
	Link Function = complementary log-log						Link Function = logistic			
Variables		Mean	SD	Z	p-value		Mean	SD	Z	p-value
Time Points	1.14	-7.86	0.42	-18.72	<0.0001		-8.72	0.5	-17.28	<0.0001
	2.02	-7.67	0.4	-18.97	<0.0001		-8.52	0.49	-17.42	0
	2.87	-7.26	0.39	-18.62	<0.0001		-8.09	0.47	-17.03	0
	3.91	-6.77	0.36	-18.56	<0.0001		-7.55	0.45	-16.89	0
	5.1	-6.48	0.37	-17.64	<0.0001		-7.25	0.45	-16.12	0
	6.5	-6.19	0.36	-17.01	<0.0001		-6.92	0.44	-15.58	0
	7.93	-5.94	0.36	-16.5	<0.0001		-6.64	0.44	-15.14	0
	9.96	-5.68	0.36	-15.8	<0.0001		-6.34	0.44	-14.54	0
	12.51	-5.44	0.36	-15.08	<0.0001		-6.07	0.44	-13.91	0
	15.66	-5.2	0.36	-14.46	<0.0001		-5.79	0.43	-13.33	0
	19.68	-4.93	0.36	-13.72	<0.0001		-5.46	0.43	-12.65	0
	24.84	-4.67	0.36	-13.09	<0.0001		-5.13	0.43	-12	0
Nodal Status	1	-0.13	0.11	-1.11	0.27		-0.16	0.14	-1.16	0.25
Tumor Size		-0.003	0.003	-0.95	0.34		-0.003	0.004	-0.71	0.48
Age		0.07	0.005	12.57	<0.0001		0.08	0.006	12.05	0
Time Points = 6										
	Link Function = complementary log-log						Link Function = logistic			
Variables		Mean	SD	Z	p-value		Mean	SD	Z	p-value
Time Points	2.02	-7.51	0.39	-19.04	<0.0001		-8.38	0.48	-17.48	<0.0001
	3.91	-6.62	0.35	-18.66	<0.0001		-7.43	0.44	-16.96	<0.0001
	6.5	-6.03	0.35	-17.13	<0.0001		-6.79	0.43	-15.67	<0.0001
	9.96	-5.52	0.35	-15.89	<0.0001		-6.22	0.43	-14.61	<0.0001
	15.66	-5.04	0.35	-14.52	<0.0001		-5.67	0.43	-13.4	<0.0001
	24.84	-4.52	0.35	-13.07	<0.0001		-5.01	0.43	-12	<0.0001
Nodal Status	1	-0.17	0.11	-1.53	0.13		-0.21	0.14	-1.56	0.12
Tumor Size		-0.003	0.003	-1.02	0.31		-0.003	0.004	-0.79	0.43
Age		0.06	0.005	12.53	<0.0001		0.08	0.006	12.05	<0.0001

<b>Table 7 Continued</b>										
Time Points = 4										
	Link Function = complementary log-log						Link Function = logistic			
Variables		Mean	SD	Z	p-value		Mean	SD	Z	p-value
Time Points	2.87	-6.87	0.36	-18.86	<0.0001		-7.69	0.45	-17.16	<0.0001
	6.5	-5.8	0.34	-17.18	<0.0001		-6.54	0.42	-15.66	<0.0001
	12.5	-5.05	0.33	-15.17	<0.0001		-5.69	0.41	-13.94	<0.0001
	24.84	-4.3	0.33	-12.99	<0.0001		-4.77	0.4	-11.89	<0.0001
Nodal Status	1	-0.21	0.11	-1.94	0.05		-0.26	0.13	-1.99	0.047
Tumor Size		-0.004	0.003	-1.22	0.22		-0.003	0.004	-0.98	0.33
Age		0.06	0.005	12.41	<0.0001		0.07	0.006	11.94	<0.0001

Comparing the Cox proportional hazards model and generalized linear model with GEE based on the pseudo-observations, all models for breast cancer-related death show that nodal status and tumor size had significant effects. However, the effect of age was different between Cox model and generalized linear model with GEE. All models for death not due to breast cancer show that the effect of nodal status, tumor size and age are the same.

## 4.0 Discussion

The pseudo observation method makes regression modeling for censored survival data more flexible, allowing for different link functions [11]. Compared to the Cox model for cause-specific hazard analysis, the generalized linear model with GEE based on pseudo observations does not require strict assumption such as the proportional hazards assumption. Because the method replaces the censoring and complete data with the pseudo values, one can use the standard generalized linear model to estimate the parameters and evaluate the relationship between covariates and cumulative incidence directly.

In this thesis, we used the independent working covariance structure. Klein and Andersen (2007) conducted a Monte Carlo study to assess whether different working covariances affect the estimation in GEE procedure [1]. They compared independent working covariance, exact working covariance and empirical working covariance and used five time points. The result showed that there was no significant difference in parameter estimates using different working covariance structure. Based on their results, we chose the independent working covariance which is the simplest working covariance structure. The results of B-04 study show that the different number of time points does not influence the estimates from the GEE procedure. Klein and Anderson (2007) also performed the Monte Carlo simulation, and the result showed that number of time points does not have influence on the model fit [1].

The results of the B-04 study seem to show that the link function does not influence the effects of covariates on cumulative incidence function. The standard deviations of covariates in the logistic link function model is larger than ones from the complementary log-log link function model.

Although the result of the B-04 study shows the link function does not effects of covariates, the choice link function introduces different model assumptions in general. According to Monte Carlo study, the different number of time points do not affect the effect of the covariates on cumulative incidence. However, Andersen and Klein suggested that 5 to 10 time points would be sufficient to provide reasonable estimates.

## Appendix A R code

### B-04 Study

#### R code for cause-specific Cox model

```
b04<-read.table("C:\\Users\\cszhe\\OneDrive\\Documents\\thesis\\pseudo
```

```
observation\\b04.with.covariates.csv",sep="," ,header = T)
```

```
b04$age <- b04$age*100
```

```
b04$tsize <- b04$tsize*100
```

```
library(survival)
```

```
library(cmprsk)
```

```
library(ggplot2)
```

```
library(zoo)
```

```
library(survminer)
```

```
#summary of event status and nodal status
```

```
table(b04$event.status)
```

```
table(b04$nodal.status)
```

```
# draw cumulative incidence curve
```

```
b04$nodal.status[b04$nodal.status==0]="negative"
```

```
b04$nodal.status[b04$nodal.status==1]="positive"
```

```
cif1 <- cuminc(ftime = b04$time,fstatus = b04$event.status)
```

```

plot(cif1,col=1:2,xlab = "Year", main = "CIC of type1 and type2 event",wh=c(0,5))

legend("topleft",c("death with breast cancer recurring","death not due to breast cancer"),
lty=c(1,2), col=c(1,2))

# cox model for type1 event

coxfit1 <- coxph(Surv(time,event.status==1)~nodal.status+age+tsize,data=b04)

summary(coxfit1)


# proportional hazard assumption check for type1

test.ph1 <- cox.zph(coxfit1)

test.ph1

ggcoxzph(test.ph1)


# cox model for type 2 event

coxfit2 <- coxph(Surv(time,event.status==2)~nodal.status+age+tsize,data=b04)

summary(coxfit2)


# proportional hazard assumption check for type2

test.ph2 <- cox.zph(coxfit2)

test.ph2

ggcoxzph(test.ph2)


R code for generalized linear model with GEE based on pseudo-observations

library(KMsurv)

```



```

library(geepack)

library(pseudo)

library(MuMIn)


b04<-read.table("C:\\Users\\cszhe\\OneDrive\\Documents\\thesis\\pseudo
observation\\b04.with.covariates.csv",sep="," ,header = T)


# choose the cut off points

sum(as.numeric(b04$event.status!=0))

b04.event=subset(b04,b04$event.status!=0)

b04.event=b04.event[order(b04.event$time),]

cutoff1=b04.event[c(100,200,300,400,500,600,700,800,900,1000,1100,1200),]$time

cutoff2=b04.event[c(200,400,600,800,1000,1200),]$time

cutoff3=b04.event[c(300,600,900,1200),]$time


# multiply 100 into age and tsize

b04$age <- b04$age*100

b04$tsize <- b04$tsize*100


# relapse

# cutoff1

```

```

# generate pseudo observations

pseudo <- pseudoci(time=b04$time,event=b04$event.status,tmax=cutoff1)


# rearrange the data into a long data set, use only pseudo-observations for type2

b <- NULL

for(it in 1:length(pseudo$time)){

  b <- rbind(b,cbind(b04,pseudo = pseudo$pseudo[[2]][it],

                    tpseudo = pseudo$time[it],id=1:nrow(b04)))

}

b <- b[order(b$id),]


# fit the model- cloglog1

library(geepack)

fit <- geese(pseudo ~ as.factor(tpseudo) + as.factor(b$nodal.status) + tsize +
            age -1, data =b, id=id, jack = TRUE, scale.fix=TRUE, family=gaussian,
            mean.link = "cloglog", corstr="independence")


#The results using the AJ variance estimate

fit.cloglog1 = cbind(mean = round(fit$beta,4), SD = round(sqrt(diag(fit$vbeta.ajs)),4),

                    Z = round(fit$beta/sqrt(diag(fit$vbeta.ajs)),4),

```

```

PVal = round(2-2*pnorm(abs(fit$beta/sqrt(diag(fit$vbeta.ajs)))),4))

fit.cloglog1

#fit the model - logit1

fit <- geese(pseudo ~ as.factor(tpseudo) + as.factor(b$nodal.status) + tsize +
            age -1, data =b, id=id, jack = TRUE, scale.fix=TRUE, family=gaussian,
            mean.link = "logit", corstr="independence")

#The results using the AJ variance estimate

fit.logit1 = cbind(mean = round(fit$beta,4), SD = round(sqrt(diag(fit$vbeta.ajs)),4),
                  Z = round(fit$beta/sqrt(diag(fit$vbeta.ajs)),4),
                  PVal = round(2-2*pnorm(abs(fit$beta/sqrt(diag(fit$vbeta.ajs)))),4))

fit.logit1

# cutoff2

# generate pseudo observations

pseudo <- pseudoci(time=b04$time,event=b04$event.status,tmax=cutoff2)

```

```

# rearrange the data into a long data set, use only pseudo-observations for type2

b <- NULL

for(it in 1:length(pseudo$time)){

  b <- rbind(b,cbind(b04,pseudo = pseudo$pseudo[[2]][,it],

                    tpseudo = pseudo$time[it],id=1:nrow(b04)))

}

b <- b[order(b$id),]


# fit the model-cloglog2

library(geepack)

fit <- geese(pseudo ~ as.factor(tpseudo) + as.factor(b$nodal.status) + tsize +

            age -1, data =b, id=id, jack = TRUE, scale.fix=TRUE, family=gaussian,

            mean.link = "cloglog", corstr="independence")


#The results using the AJ variance estimate

fit.cloglog2 = cbind(mean = round(fit$beta,4), SD = round(sqrt(diag(fit$vbeta.ajs)),4),

                    Z = round(fit$beta/sqrt(diag(fit$vbeta.ajs)),4),

                    PVal = round(2-2*pnorm(abs(fit$beta/sqrt(diag(fit$vbeta.ajs))))),4))

fit.cloglog2

```

```

# fit the model-logit2

library(geepack)

fit <- geese(pseudo ~ as.factor(tpseudo) + as.factor(b$nodal.status) + tsize +
            age -1, data =b, id=id, jack = TRUE, scale.fix=TRUE, family=gaussian,
            mean.link = "logit", corstr="independence")


#The results using the AJ variance estimate

fit.logit2 = cbind(mean = round(fit$beta,4), SD = round(sqrt(diag(fit$vbeta.ajs)),4),
                  Z = round(fit$beta/sqrt(diag(fit$vbeta.ajs)),4),
                  PVal = round(2-2*pnorm(abs(fit$beta/sqrt(diag(fit$vbeta.ajs))))),4))

fit.logit2


# cutoff3

# generate pseudo observations

pseudo <- pseudoci(time=b04$time,event=b04$event.status,tmax=cutoff3)


# rearrange the data into a long data set, use only pseudo-observations for type2

b <- NULL

for(it in 1:length(pseudo$time)){

  b <- rbind(b,cbind(b04,pseudo = pseudo$pseudo[[2]][it],

```

```

        tpseudo = pseudo$time[it],id=1:nrow(b04)))
    }
b <- b[order(b$id),]

# fit the model - cloglog3
library(geepack)
fit <- geese(pseudo ~ as.factor(tpseudo) + as.factor(b$nodal.status) + tsize +
            age -1, data =b, id=id, jack = TRUE, scale.fix=TRUE, family=gaussian,
            mean.link = "cloglog", corstr="independence")

#The results using the AJ variance estimate
fit.cloglog3 = cbind(mean = round(fit$beta,4), SD = round(sqrt(diag(fit$vbeta.ajs)),4),
                    Z = round(fit$beta/sqrt(diag(fit$vbeta.ajs)),4),
                    PVal = round(2-2*pnorm(abs(fit$beta/sqrt(diag(fit$vbeta.ajs))))),4))
fit.cloglog3

# fit the model - logit3
library(geepack)
fit <- geese(pseudo ~ as.factor(tpseudo) + as.factor(b$nodal.status) + tsize +
            age -1, data =b, id=id, jack = TRUE, scale.fix=TRUE, family=gaussian,
            mean.link = "logit", corstr="independence")

```

```

#The results using the AJ variance estimate

fit.logit3 = cbind(mean = round(fit$beta,4), SD = round(sqrt(diag(fit$vbeta.ajs)),4),

                  Z = round(fit$beta/sqrt(diag(fit$vbeta.ajs)),4),

                  PVal = round(2-2*pnorm(abs(fit$beta/sqrt(diag(fit$vbeta.ajs))))),4))

fit.logit3


# death

# cutoff1

# generate pseudo observations

pseudo <- pseudoci(time=b04$time,event=b04$event.status,tmax=cutoff1)


# rearrange the data into a long data set, use only pseudo-observations for type2

b <- NULL

for(it in 1:length(pseudo$time)){

  b <- rbind(b,cbind(b04,pseudo = pseudo$pseudo[[1]][,it],

                    tpseudo = pseudo$time[it],id=1:nrow(b04)))

}

b <- b[order(b$id),]


# fit the model- cloglog1

library(geepack)

fit <- geese(pseudo ~ as.factor(tpseudo) + as.factor(b$nodal.status) + tsize +

```

```

age -1, data =b, id=id, jack = TRUE, scale.fix=TRUE, family=gaussian,
mean.link = "cloglog", corstr="independence")

```

#The results using the AJ variance estimate

```

fit.cloglog1 = cbind(mean = round(fit$beta,4), SD = round(sqrt(diag(fit$vbeta.ajs)),4),
                      Z = round(fit$beta/sqrt(diag(fit$vbeta.ajs)),4),
                      PVal = round(2-2*pnorm(abs(fit$beta/sqrt(diag(fit$vbeta.ajs))))),4))
fit.cloglog1

```

#fit the model - logit1

```

fit <- geese(pseudo ~ as.factor(tpseudo) + as.factor(b$nodal.status) + tsize +
age -1, data =b, id=id, jack = TRUE, scale.fix=TRUE, family=gaussian,
mean.link = "logit", corstr="independence")

```

#The results using the AJ variance estimate

```

fit.logit1 = cbind(mean = round(fit$beta,4), SD = round(sqrt(diag(fit$vbeta.ajs)),4),
                    Z = round(fit$beta/sqrt(diag(fit$vbeta.ajs)),4),
                    PVal = round(2-2*pnorm(abs(fit$beta/sqrt(diag(fit$vbeta.ajs))))),4))
fit.logit1

```



```

# cutoff2

# generate pseudo observations

pseudo <- pseudoci(time=b04$time,event=b04$event.status,tmax=cutoff2)


# rearrange the data into a long data set, use only pseudo-observations for type2

b <- NULL

for(it in 1:length(pseudo$time)){

  b <- rbind(b,cbind(b04,pseudo = pseudo$pseudo[[1]][it],

                    tpseudo = pseudo$time[it],id=1:nrow(b04)))

}

b <- b[order(b$id),]


# fit the model-cloglog2

library(geepack)

fit <- geese(pseudo ~ as.factor(tpseudo) + as.factor(b$nodal.status) + tsize +

            age -1, data =b, id=id, jack = TRUE, scale.fix=TRUE, family=gaussian,

            mean.link = "cloglog", corstr="independence")


#The results using the AJ variance estimate

fit.cloglog2 = cbind(mean = round(fit$beta,4), SD = round(sqrt(diag(fit$vbeta.ajs)),4),

                    Z = round(fit$beta/sqrt(diag(fit$vbeta.ajs)),4),

```

```

PVal = round(2-2*pnorm(abs(fit$beta/sqrt(diag(fit$vbeta.ajs)))),4))

fit.cloglog2

# fit the model-logit2

library(geepack)

fit <- geese(pseudo ~ as.factor(tpseudo) + as.factor(b$nodal.status) + tsize +
            age -1, data =b, id=id, jack = TRUE, scale.fix=TRUE, family=gaussian,
            mean.link = "logit", corstr="independence")

#The results using the AJ variance estimate

fit.logit2 = cbind(mean = round(fit$beta,4), SD = round(sqrt(diag(fit$vbeta.ajs)),4),
                  Z = round(fit$beta/sqrt(diag(fit$vbeta.ajs)),4),
                  PVal = round(2-2*pnorm(abs(fit$beta/sqrt(diag(fit$vbeta.ajs)))),4))

fit.logit2

# cutoff3

# generate pseudo observations

pseudo <- pseudoci(time=b04$time,event=b04$event.status,tmax=cutoff3)

# rearrange the data into a long data set, use only pseudo-observations for type2

b <- NULL

for(it in 1:length(pseudo$time)){

```

```

b <- rbind(b,cbind(b04,pseudo = pseudo$pseudo[[1]][,it],
                  tpseudo = pseudo$time[it],id=1:nrow(b04)))
}

b <- b[order(b$id),]

# fit the model - cloglog3

library(geepack)

fit <- geese(pseudo ~ as.factor(tpseudo) + as.factor(b$nodal.status) + tsize +
            age -1, data =b, id=id, jack = TRUE, scale.fix=TRUE, family=gaussian,
            mean.link = "cloglog", corstr="independence")

#The results using the AJ variance estimate

fit.cloglog3 = cbind(mean = round(fit$beta,4), SD = round(sqrt(diag(fit$vbeta.ajs)),4),
                    Z = round(fit$beta/sqrt(diag(fit$vbeta.ajs)),4),
                    PVal = round(2-2*pnorm(abs(fit$beta/sqrt(diag(fit$vbeta.ajs))))),4))

fit.cloglog3

# fit the model - logit3

library(geepack)

fit <- geese(pseudo ~ as.factor(tpseudo) + as.factor(b$nodal.status) + tsize +
            age -1, data =b, id=id, jack = TRUE, scale.fix=TRUE, family=gaussian,

```

```
mean.link = "logit", corstr="independence")
```

```
#The results using the AJ variance estimate
```

```
fit.logit3 = cbind(mean = round(fit$beta,4), SD = round(sqrt(diag(fit$vbeta.ajs)),4),
```

```
  Z = round(fit$beta/sqrt(diag(fit$vbeta.ajs)),4),
```

```
  PVal = round(2-2*pnorm(abs(fit$beta/sqrt(diag(fit$vbeta.ajs))))),4))
```

```
fit.logit3
```

## Bibliography

1. Andersen, Per K., and John P. Klein. "Regression Analysis for Multistate Models Based on a Pseudo-Value Approach, with Applications to Bone Marrow Transplantation Studies." *Scandinavian Journal of Statistics*, vol. 34, no. 1, 2007, pp. 3-16.
2. Andersen, Per K., and Maja Pohar Perme. "Pseudo-Observations in Survival Analysis." *Statistical Methods in Medical Research*, vol. 19, no. 1, 2010, pp. 71-99.
3. Andersen, Per K., John P. Klein, and Susanne Rosthøj. "Generalised Linear Models for Correlated Pseudo-Observations, with Applications to Multi-State Models." *Biometrika*, vol. 90, no. 1, 2003, pp. 15-27.
4. Black, Dalliah M., MD, and Mittendorf, Elizabeth A., MD, PhD. "Landmark Trials Affecting the Surgical Management of Invasive Breast Cancer." *Surgical Clinics of North America*, the, vol. 93, no. 2, 2013, pp. 501-518
5. Beadle, Beth M., MD, PhD, Woodward, Wendy A., MD, PhD, and Thomas A. Buchholz MD. "The Impact of Age on Outcome in Early-Stage Breast Cancer." *Seminars in Radiation Oncology*, vol. 21, no. 1, 2011, pp. 26-34.
6. "Breast Cancer - Stages." *Cancer.Net*, 19 Nov. 2018, [www.cancer.net/cancer-types/breast-cancer/stages](http://www.cancer.net/cancer-types/breast-cancer/stages).
7. Ewa Wycinka, Tomasz Jurkiewicz. "Survival Regression Models for Single Events and Competing Risks Based On Pseudo-Observations." *Statistics in Transition new series*, vol. 20, no. 1, 2019, pp. 433-452.
8. Haller, Bernhard, Georg Schmidt, and Kurt Ulm. "Applying Competing Risks Regression Models: An Overview." *Lifetime Data Analysis*, vol. 19, no. 1, 2013;2012;, pp. 33-58.
9. Harbeck, Nadia, and Michael Gnant. "Breast Cancer." *The Lancet*, vol. 389, no. 10074, 2017, pp. 1134-1150.
10. Klein, John P., et al. "SAS and R Functions to Compute Pseudo-Values for Censored Data Regression." *Computer Methods and Programs in Biomedicine*, vol. 89, no. 3, 2007;2008;, pp. 289-300.
11. Klein, John P., and Per K. Andersen. "Regression Modeling of Competing Risks Data Based on Pseudovalues of the Cumulative Incidence Function." *Biometrics*, vol. 61, no. 1, 2005, pp. 223-229.
- 12 "Learn About Lymph Node Status and Breast Cancer at Susan G. Komen." Susan G. Komen®, [ww5.komen.org/BreastCancer/LymphNodeStatus.html](http://ww5.komen.org/BreastCancer/LymphNodeStatus.html).

13. McIntosh, Avery. "The Jackknife Estimation Method.", 2016.
14. Orsaria, Paolo, et al. "Nodal Status Assessment in Breast Cancer: Strategies of Clinical Grounds and Quality of Life Implications." *International Journal of Breast Cancer*, vol. 2014, 2014, pp. 469803-8.
15. Putter, H., et al. "Tutorial in Biostatistics: Competing Risks and Multi-State Models." *Statistics in Medicine*, vol. 26, no. 11, 2007, pp. 2389–2430., doi:10.1002/sim.2712.
16. Whitlock, Jennifer, and Msn. "The 3 Most Common Kinds of Mastectomies." *Verywell Health*, Verywellhealth, 23 Aug. 2017, [www.verywellhealth.com/types-of-mastectomy-breast-surgery-3157281](http://www.verywellhealth.com/types-of-mastectomy-breast-surgery-3157281).